AEM: Problem Set 1

Maxwell Austensen

September 24, 2016

```r
# Utility functions -------------------------------------------------------

`%S%` <- function(x, y) {
  paste0(x, y)
}

`%notin%` <- Negate(`%in%`)

################################################################################

# Install packages if needed
package_list <- c("stargazer", "knitr", "haven", "labelled", "ICC", "scales", "tidyverse")
new_packages <- package_list[package_list %notin% installed.packages()[,"Package"]]
if(length(new_packages)) install.packages(new_packages)

library(stargazer)
library(knitr)
library(haven)
library(labelled)
library(ICC)
library(scales)
library(tidyverse)

# Set directories
repo_ <- "H:/GitHub/aem/"
ps1_ <- "C:/Users/austensen/Box Sync/aem/ps1/"

################################################################################
```

```r
# Load data
data_raw <-
  read_stata(ps1_ %S% "Thornton HIV Testing Data.dta") %>%
  remove_val_labels

names(data_raw) <- names(data_raw) %>% tolower

# Contruct main sample for analysis
main_sample <-
  data_raw %>%
  filter(
    hiv2004 %notin% c(NA, -1),
    !is.na(any),
    !is.na(zone),
    !is.na(age)
  )
```

Part I: Summary Statistics

```
get_summary <- function(data){
  data %>%
    summarise(
      `Average Age` = mean(age, na.rm=T),
      `Percentage of Males` = mean(male, na.rm=T)*100,
      `Average Years of Education` = mean(educ2004, na.rm=T),
      `Percentage with HIV` = mean(hiv2004, na.rm=T)*100
    ) %>%
    kable(digits = 1)
}
```

1.

```
main_sample %>% get_summary
```

Average Age	Percentage of Males	Average Years of Education	Percentage with HIV
33.4	46.3	3.6	6.3

The average age in the sample is 33.4 years old, the sample is 46.3% males, and 6.3 percent of people in the sample are infected with HIV.

2.

```
main_sample %>% group_by(any) %>% get_summary
```

any	Average Age	Percentage of Males	Average Years of Education	Percentage with HIV
0	32.1	47.1	4.5	6.3
1	33.7	46.1	3.4	6.2

```
main_sample %>% group_by(under) %>% get_summary
```

under	Average Age	Percentage of Males	Average Years of Education	Percentage with HIV
0	33.7	46.5	3.5	5
1	33.0	46.0	3.9	8

There are no major differences in the variables between treatment and control groups based on either cash receipt or distance. However, those that received some cash were on average about a year older and those that were under 1.5Km from the center had an HIV rate three percentage points high than those further away.

3.

```r
grps <- c("any", "under")
vars <- c("age", "male", "hiv2004")

ttest_results <-
  data.frame(
    var = character(),
    P.value = double(),
    stringsAsFactors=FALSE
  )

for(grp in grps) {
  i <- 1
  for(var in vars){
    result <- t.test(main_sample[[var]] ~ main_sample[[grp]], var.equal = TRUE)$p.value

    ttest_results[i,1] <- var
    ttest_results[i,2] <- result
    i <- i+1
  }
  writeLines("#### t-test: group = " %S% grp)
  print(kable(ttest_results, digits = 3))
  writeLines("\n\n\n")
}
```

t-test: group = any

var	P.value
age	0.008
male	0.658
hiv2004	0.952

t-test: group = under

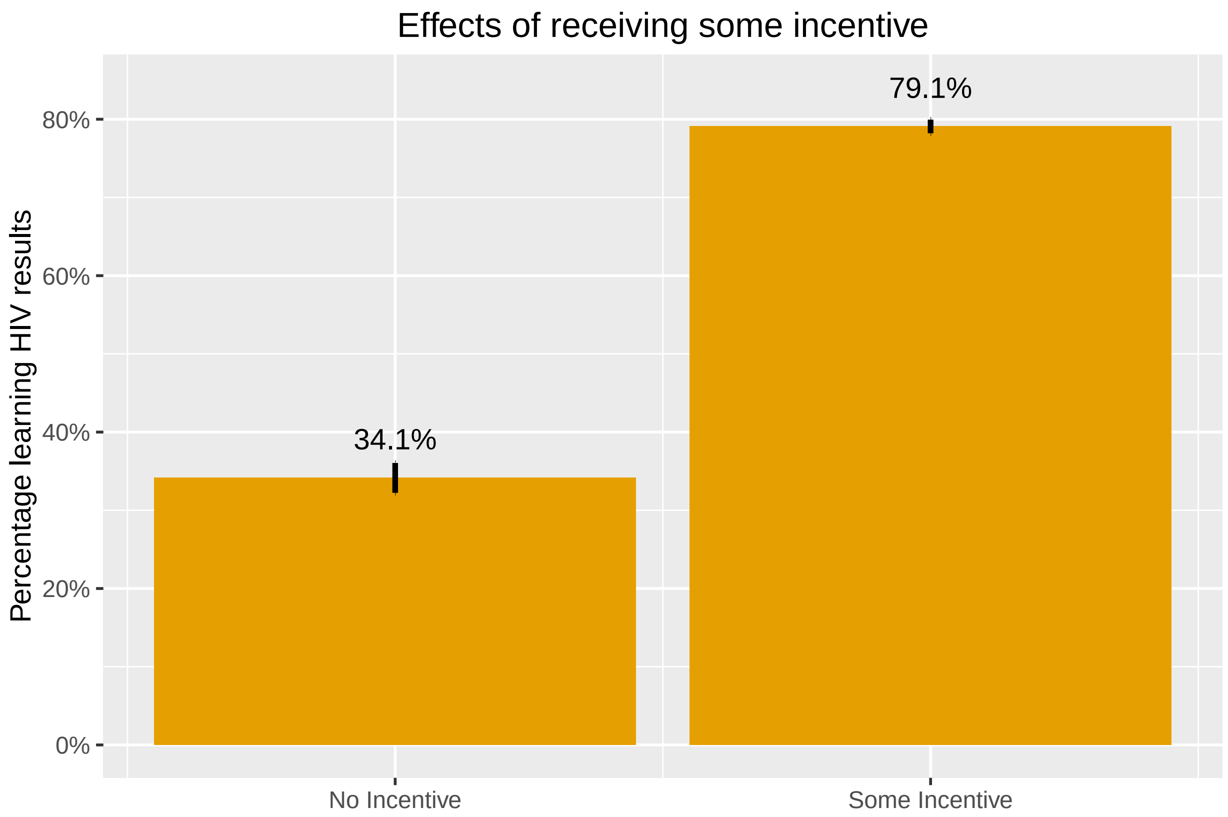
var	P.value
age	0.194
male	0.813
hiv2004	0.001

Those in the sample that received any cash were older than those that did not receive any cash, and this difference is significant at the 1% level. Also, Those in the sample that were less than 1.5Km from the center were more likely to be infected with HIV than those further from the centers, and this difference was also significant at the 1% level.

Part II: Analysis using graphs

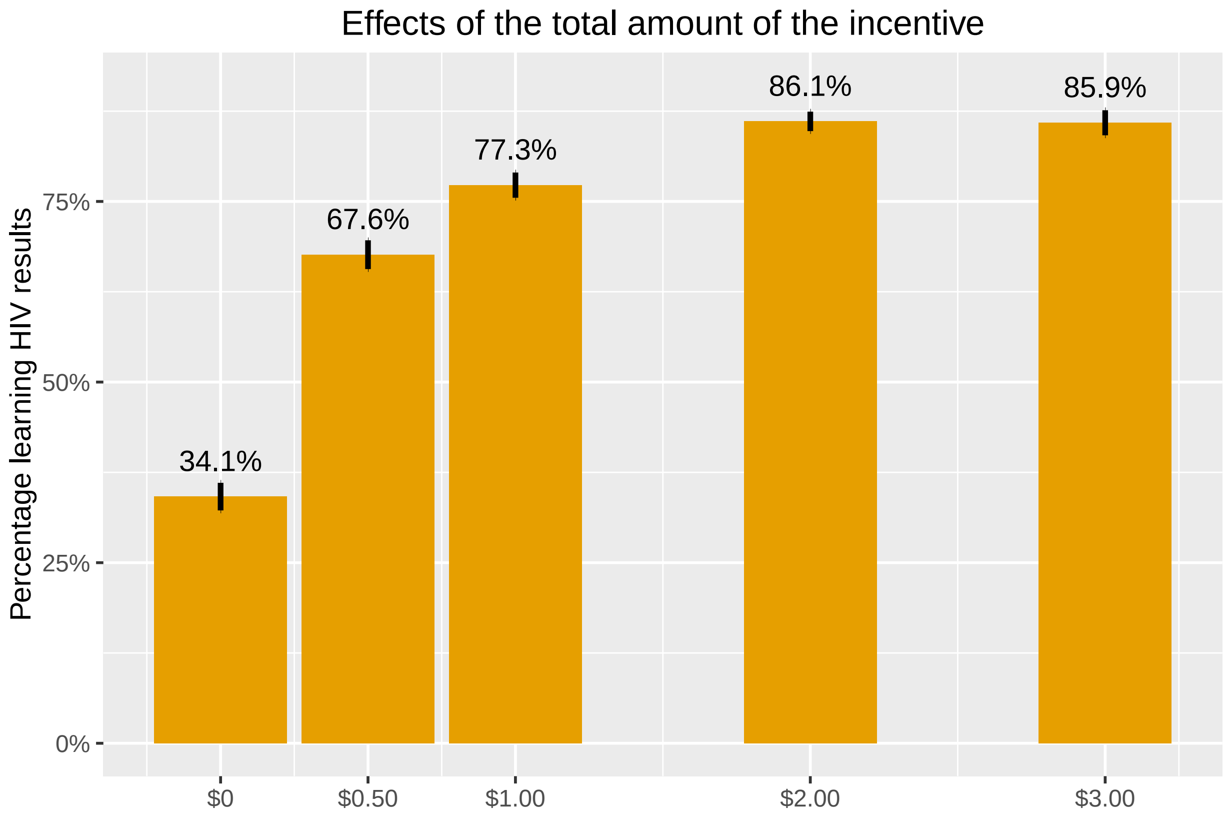
4.

```
main_sample %>%
  group_by(any) %>%
  summarise(
    got_mean = mean(got, na.rm=T),
    got_se = sd(got)/sqrt(n())
  ) %>%
  ggplot(aes(any, got_mean)) +
  geom_bar(stat = "identity", fill = "#E69F00") +
  geom_errorbar(aes(ymin = got_mean - got_se, ymax = got_mean + got_se), width = 0, size = 1) +
  geom_text(aes(y = got_mean+0.05, label = round(got_mean*100, 1) %S% "%")) +
  scale_y_continuous(labels = percent_format()) +
  scale_x_continuous(breaks = c(0,1), labels = c("No Incentive", "Some Incentive")) +
  ylab("Percentage learning HIV results") + xlab("") +
  ggtitle("Effects of receiving some incentive")
```



5.

```
main_sample %>%
  group_by(ti) %>%
  summarise(
    got_mean = mean(got, na.rm=T),
    got_se = sd(got)/sqrt(n())
  ) %>%
  ggplot(aes(ti, got_mean)) +
  geom_bar(stat = "identity", fill = "#E69F00") +
  geom_errorbar(aes(ymin = got_mean - got_se, ymax = got_mean + got_se), width = 0, size = 1) +
  geom_text(aes(y = got_mean+0.05, label = round(got_mean*100, 1) %S% "%")) +
  scale_y_continuous(labels = percent_format()) +
  scale_x_continuous(breaks = c(0,50, 100, 200, 300), labels = c("$0", "$0.50", "$1.00", "$2.00", "$3.0
  ylab("Percentage learning HIV results") + xlab("") +
  ggtitle("Effects of the total amount of the incentive")
```



Part III: Analysis using linear regression

6.

```r
any_1 <- lm(got ~ any, data = main_sample)
any_2 <- lm(got ~ any + age + male + educ2004 + mar, data = main_sample)

stargazer(any_1, any_2, type = output_type, header = FALSE)
```

Table 6:

	Dependent variable:	
	got	
	(1)	(2)
any	0.449***	0.450***
	(0.019)	(0.020)
age		0.001
		(0.001)
male		−0.010
		(0.017)
educ2004		−0.009***
		(0.003)
mar		0.013
		(0.022)
Constant	0.341***	0.362***
	(0.017)	(0.039)
Observations	2,812	2,530
R^2	0.162	0.181
Adjusted R^2	0.162	0.179
Residual Std. Error	0.423 (df = 2810)	0.412 (df = 2524)
F Statistic	545.100*** (df = 1; 2810)	111.239*** (df = 5; 2524)
Note:		*p<0.1; **p<0.05; ***p<0.01

The estimate for b is 0.449, and it is statistically significant at the 1% level. When additional controls are included the estimate is virtually unchanged at 0.450, and remains significant at the 1% level. This suggests that the randomization was successful in balancing the treatment and control groups, and suggests that there is no covert or overt bias in the treatment effect estimate.

7.

```r
results <- t.test(main_sample$got ~ main_sample$any, var.equal=TRUE, paired=FALSE)

results$p.value
```

```
## [1] 2.440813e-110
```

```r
results$estimate[[2]] - results$estimate[[1]]
```

```
## [1] 0.4493691
```

Using a group means comparison, the estimated treatment effect is 0.449, and is statistically significant at the 1% level. This answer does not differ significantly from the OLS coefficient estimate on treatment. Since the estimate using regressions with control variables included does not significantly alter the treatment effect estimate from a simple mean difference, this suggests that the randomization of treatment was successful in balancing the two groups with respect to these other variables.

8.

```r
ti_1 <- lm(got ~ ti, data = main_sample)
ti_2 <- lm(got ~ ti + age + male + educ2004 + mar, data = main_sample)

stargazer(ti_1, ti_2, type = output_type, header = FALSE)
```

Table 7:

	Dependent variable:	
	got	
	(1)	(2)
ti	0.002***	0.002***
	(0.0001)	(0.0001)
age		0.001
		(0.001)
male		−0.024
		(0.018)
educ2004		−0.013***
		(0.003)
mar		0.003
		(0.022)
Constant	0.499***	0.551***
	(0.013)	(0.037)
Observations	2,812	2,530
R^2	0.125	0.142
Adjusted R^2	0.125	0.141
Residual Std. Error	0.432 (df = 2810)	0.422 (df = 2524)
F Statistic	401.536*** (df = 1; 2810)	83.879*** (df = 5; 2524)
Note:		*p<0.1; **p<0.05; ***p<0.01

The estimate of the treatment effect is 0.002, meaning that, on average, an additional $1.00 in cash incentive is associated with a 2% increase in the likelihood of getting the test results. The addition of control variables does not effect the estimate of the treatment effect at all.

9.

Receiving any cash incentive increases the likelihood of the recipient getting their HIV test results by 50%. This is quite a large effect. Doubling the cash incentive from $1 to $2 has a relatively small effect compared to that from moving from no incentive to $0.50.

Part IV: Conditional (Heterogeneous) Treatment Effects

10.

```r
any_male <- lm(got ~ any + male + any*male, data = main_sample)

stargazer(any_male, type = output_type, header = FALSE)
```

Table 8:

	Dependent variable:
	got
any	0.445***
	(0.026)
male	−0.015
	(0.034)
any:male	0.009
	(0.039)
Constant	0.349***
	(0.023)
Observations	2,812
R^2	0.163
Adjusted R^2	0.162
Residual Std. Error	0.423 (df = 2808)
F Statistic	181.701*** (df = 3; 2808)
Note:	*p<0.1; **p<0.05; ***p<0.01

The estimate for the treatment-male interaction term is -0.015, and it is not statistically significant at the 10% level. This suggests that there is not a differential effect of receiving any cash incentive for men and women. The estimate would be interpreted as meaning that men who receive any cash incentive are 0.6% less likely to get their results than are women who receive any cash incentive. The interpretation of the coefficient for the interaction `male*any` is different than for `male` because the interaction must be interpreted in combination with the coefficients on its constituent parts.

11.

```
any_educ <- lm(got ~ any + educ2004 + any*educ2004, data = main_sample)

stargazer(any_educ, type = output_type, header = FALSE)
```

Table 9:

	Dependent variable:
	got
any	0.446***
	(0.030)
educ2004	−0.010**
	(0.005)
any:educ2004	0.001
	(0.005)
Constant	0.394***
	(0.027)
Observations	2,530
R^2	0.180
Adjusted R^2	0.179
Residual Std. Error	0.412 (df = 2526)
F Statistic	184.730*** (df = 3; 2526)
Note:	*p<0.1; **p<0.05; ***p<0.01

The estimate of the coefficient on the interaction `any*educ2004` is 0.001, and it is not statistically significant. The interaction suggests that for those who receive any cash incentive, each additional year of education is associated with being 0.99% less likely to get the test results.

Part V: Policy Implications

12.

Based on the findings in Part III, if the goal of the government were to increase the number of people who know their HIV status, they should pursue a policy of granting some amount of cash incentive. Assuming that the cost of administering an additional test is significantly more than $2.00, the cash incentives offered should be $2.00.

13.

Given that results of Part IV show that there are no significant differential effects for the cash incentives on the likelihood of getting the results, there should not be any targeting of certain groups for the incentives.

Part VI: A Random Sub-Sample

14.

```r
set.seed(19920428)
sample_1000 <- main_sample %>% sample_n(1000)
```

15.

```r
any_1 <- lm(got ~ any, data = sample_1000)
any_2 <- lm(got ~ any + age + male + educ2004 + mar, data = sample_1000)

stargazer(any_1, any_2, type = output_type, header = FALSE)
```

The estimate of the coefficient on treatment is slightly different in this random sub sample. However, it is quite close to the previous estimate. It is different because of random variation from the sampling mean.

Part VII: Choosing Sample Size

16.

```r
sample_size <- function(mu_diff, sds, kappa, alpha, beta){
  pooled_sd <- sqrt((sds[[1]]^2 + sds[[2]]^2)/2)
  nB <- (1+1/kappa) * (pooled_sd*(qnorm(1-alpha/2) + qnorm(1-beta))/mu_diff)^2
  N <- ceiling(nB)*2
  return(N)
}

condom_sds <-
  main_sample %>%
  group_by(any) %>%
  summarise(sd = sd(numcond, na.rm = TRUE))

sd1 <- condom_sds[[1, 2]]
sd2 <- condom_sds[[2, 2]]

sample_size(mu_diff = 1, sd = c(sd1, sd2), kappa = 1, alpha = 0.05, beta = 0.20)
```

Table 10:

	Dependent variable:	
	got	
	(1)	(2)
any	0.485^{***}	0.487^{***}
	(0.032)	(0.034)
age		0.00000
		(0.001)
male		-0.007
		(0.029)
educ2004		-0.003
		(0.005)
mar		0.011
		(0.036)
Constant	0.317^{***}	0.333^{***}
	(0.028)	(0.067)
Observations	1,000	899
R^2	0.190	0.200
Adjusted R^2	0.189	0.195
Residual Std. Error	0.414 (df = 998)	0.409 (df = 893)
F Statistic	233.703^{***} (df = 1; 998)	44.535^{***} (df = 5; 893)
Note:		*p<0.1; **p<0.05; ***p<0.01

```
## [1] 114
```

```
sample_size(mu_diff = 1, sd = c(sd1, sd2), kappa = 1, alpha = 0.05, beta = 0.10)
```

```
## [1] 152
```

With power level as 0.8 a sample of 114 people will be required, and for power of 0.9 the sample will need to include 152 people.

17.

```
(This question answered in STATA)

sum numcond if any==0
sum numcond if any==1

loneway numcond site

sampsi 0 1, sd1(1.884872) sd2(1.917499) power(0.8) alpha(0.05)
sampclus, obsclus(40) rho(0.07897)

sampsi 0 1, sd1(1.884872) sd2(1.917499) power(0.9) alpha(0.05)
sampclus, obsclus(40) rho(0.07897)
```

With power of 0.8 the minimum number of clusters required would be 12, and with power of 0.9 16 clusters would be needed.

Part VIII: Fisher Randomization Test (bonus)

18.

```
fisher_test <- function(data, test_diff, reps){
  test_diffs <- vector("list", 1000)
  sim_diffs <- vector("list", 1000)

  for(i in 1:reps){

    df <- data %>% mutate(trt = sample(0:1, n(), replace = TRUE))

    ttest <- t.test(df$got ~ df$trt, var.equal=TRUE, paired=FALSE)

    sim_diff <- ttest$estimate[[2]] - ttest$estimate[[1]]

    sim_diffs[[i]] <- sim_diff
    test_diffs[[i]] <- ifelse(abs(sim_diff) > test_diff, 1, 0)
  }

  output <-
    tibble(
      sim_diffs = flatten_dbl(sim_diffs),
      test_diffs = flatten_dbl(test_diffs)
    )

  return(output)
}


ttest <- t.test(main_sample$got ~ main_sample$any, var.equal=TRUE, paired=FALSE)

any_diff <- ttest$estimate[[2]] - ttest$estimate[[1]]

data <- main_sample %>% select(got)

set.seed(19920428)

(result1 <-
  data %>%
  fisher_test(any_diff, 1000) %>%
  summarise(mean(test_diffs)) %>%
  .[[1]])
```

```
## [1] 0
```

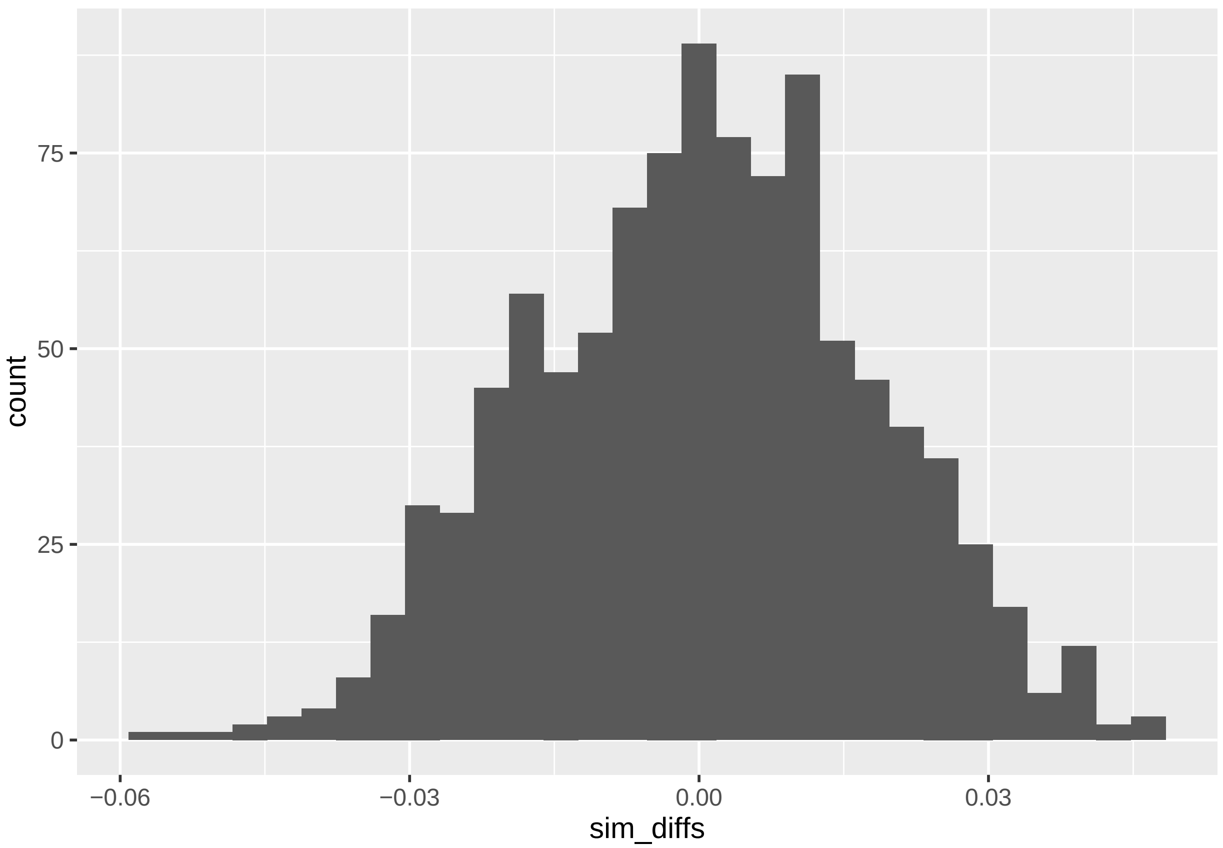
```
(result2 <-
data %>%
  fisher_test(0.05, 1000) %>%
  summarise(mean(test_diffs)) %>%
  .[[1]])
```

```
## [1] 0.007
```

```
(result3 <-
data %>%
  fisher_test(0.01, 1000) %>%
  summarise(mean(test_diffs)) %>%
  .[[1]])
```

```
## [1] 0.561
```

```
data %>%
  fisher_test(any_diff, 1000) %>%
  ggplot(aes(sim_diffs)) +
  geom_histogram()
```



Using 1,000 simulations, the simulated probability of observing a mean difference greater than the difference in getting HIV test results between those who did and did not receive any cash incentive that we in fact observe in the data is 0. The simulated probability of observing differences greater in absolute value than 0.05 and 0.01 is, respectively, 0.007 and 0.561.